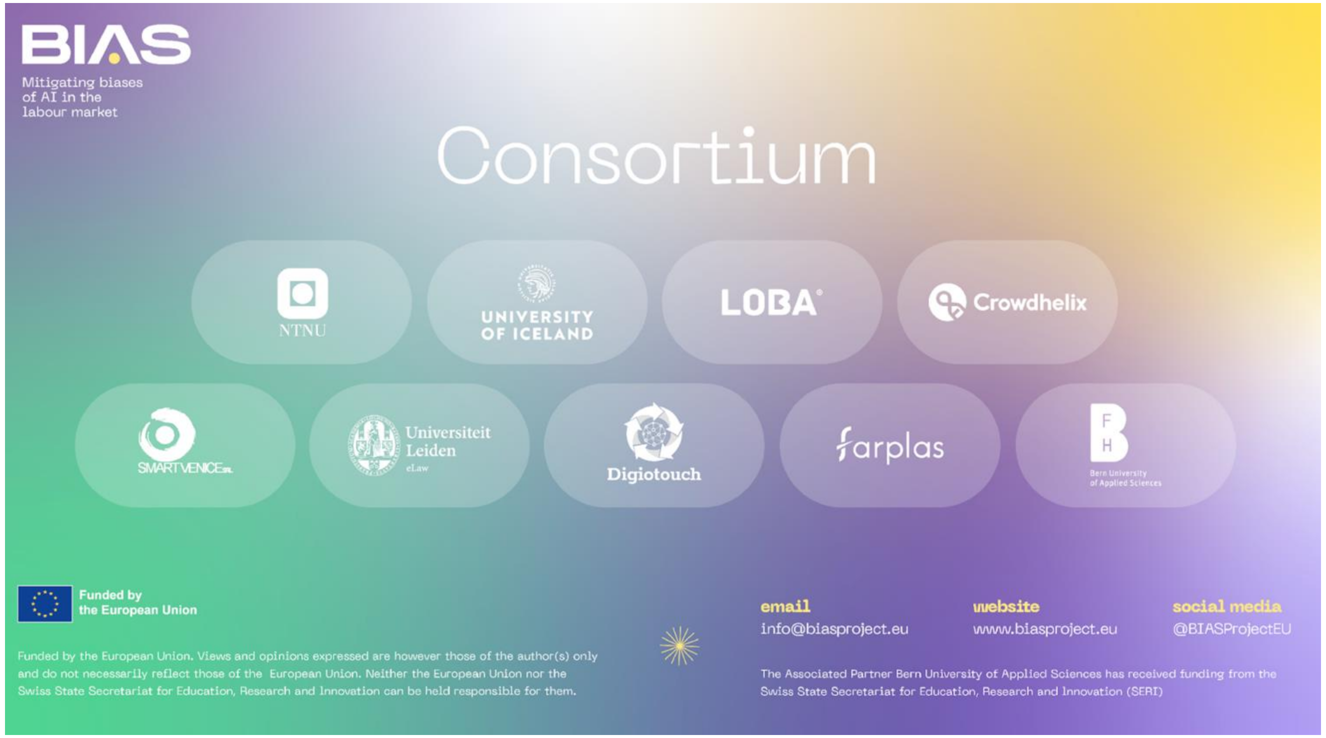
# data cyborgs

Constructing Datafied Relations, Week 9

# The BIAS Project

- Nov. 2022- Nov. 2026
- EU Horizon Project
- Identifying and mitigating diversity biases of AI applications in the labor market
- 9 partner countries
- *Interdisciplinary*: STS, law, computer science, industry, outreach/networking, commercialization



# Prompt: What is the definition of diversity bias?

*ChatGPT*: Diversity bias refers to a type of bias that occurs when efforts to promote diversity lead to unintended discrimination, favoritism, or misrepresentation. It can manifest in several ways, such as…

Tokenism – When organizations include individuals from diverse backgrounds for appearance’s sake rather than fostering genuine inclusion.

Reverse Discrimination – When attempts to promote diversity result in bias against traditionally dominant groups.

Diversity bias can occur in hiring, media representation, education, and decision-making processes…

***How many AI tools for HR in the market?***

**Over 250**


# VSD, Candidate Recommender Systems and Fairness

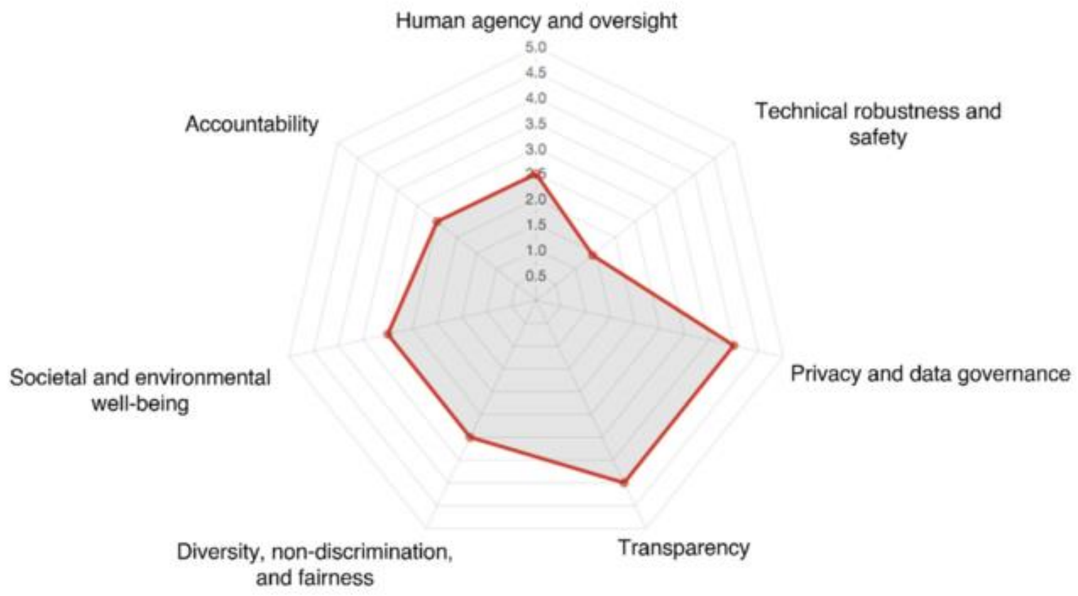
What are the values and assumptions underlying design choices for existing systems?

How do they relate to bias?

# Values and Stakeholders

Values
- Diversity, Non-Discrimination, Fairness
- Bias Detection and Mitigation
- Remaining ALTAI values

Stakeholders
- **Employers**
- **Job seekers**
- Tool developers
- Activist/Advocacy groups
- Employees
- Policy makers/Regulators



# Conceptual Investigation

- Understanding *bias* and *fairness* in recruitment/AI contexts
- Relevant policy frameworks, e.g., The *AI Act* and *EU non-discrimination law*
- Potential conflicts/trade-offs

# Conceptual Investigation

EU law

- Non-discrimination law protects against *direct* and *indirect discrimination* (latter applied to *proxy discrimination* in AI cases)
- AI Act places recruitment in *high-risk tier*.
    - *Requirements*: risk management, data governance, compliance docs, logs, transparency to users about limitations/capabilities, human oversight…

Job seekers

- *job-relatedness, consistent procedure, opportunity to perform*.

HR professionals

- *Most capable candidate, diversity balanced with "company culture"*

# Empirical Investigation

Surveys

- Ca. 1000 respondents across EU
- Multiple-choice Qs about interaction with AI in labor market and workplace

Field work/Interviews

- Expert interviews representing different stakeholders
- Ethnographic visits and interviews with companies and employees using AI tools for recruitment/work management

# Empirical Investigation

Co-creation Workshops

- In all 9 countries, careful representation across age, gender, sexual orientation, disability, culture…
- Understanding stakeholder perspectives on fairness
- Co-design VSD workshops
- Use of mock tools to stimulate discussion

Aggregating social media data

- FINDHR sister-project aggregates and taxonomizes job-seekers fairness concerns based on Reddit posts.

# Empirical Investigation

Industry collaboration

- Data collected from *FARPLAS*, where tool prototype will be deployed.
- Extensive dialogue with HR team, thorough understanding of hiring procedure.
- Understanding of potential bias in particular domain/cultural context wrt recruitment.

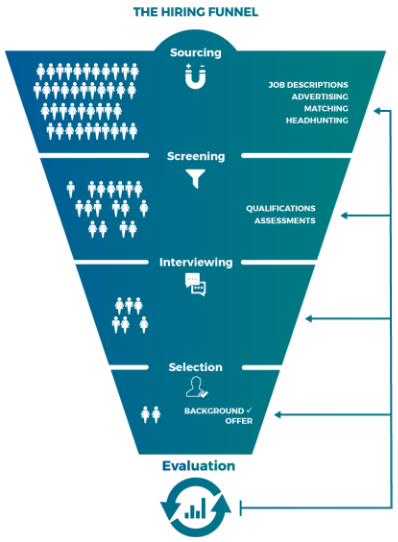
# Empirical Investigation - Findings on fairness

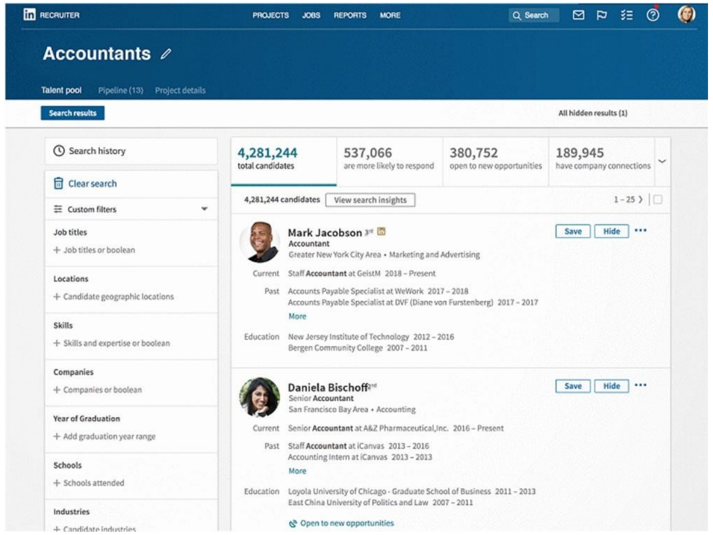
- More agency and transparency for both users and candidates
- Unfair if decisions are not justifiable
- Technical robustness is key for fairness
- Concern over unfair treatment of outlier/unique profiles
- Fear that scaled up algorithm leads to scaled up systemic bias
  - Vs. diversity of human evaluators compensating individual bias

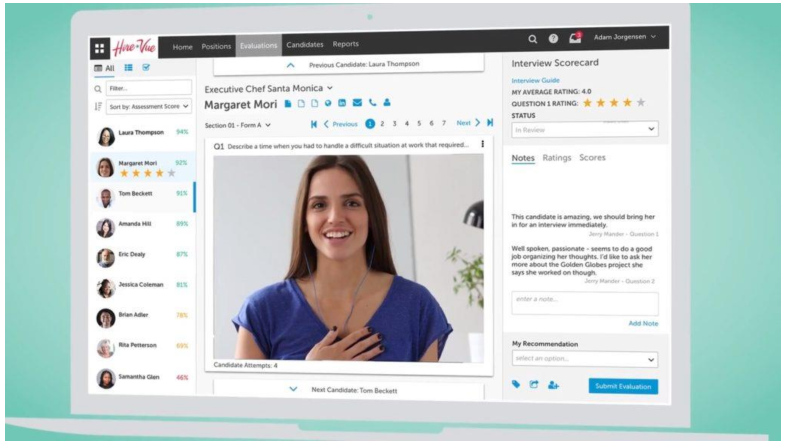
# Technical Investigation – Retrospective Analysis

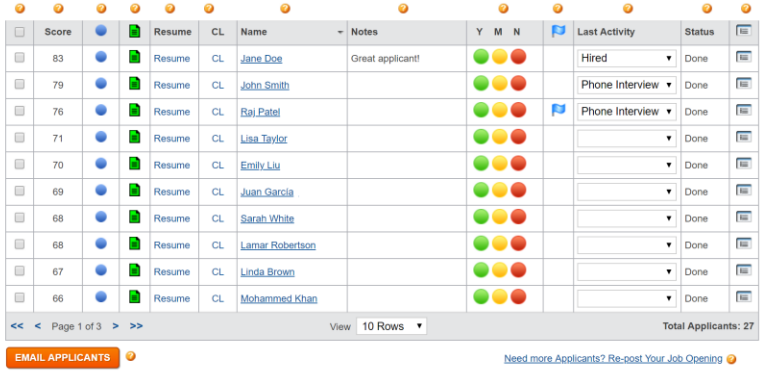
# VSD Perspective on Typical AI Candidate Ranking Systems

*Case Study:* Blackbox AI-Tools that process candidate data and output ranked recommendations.

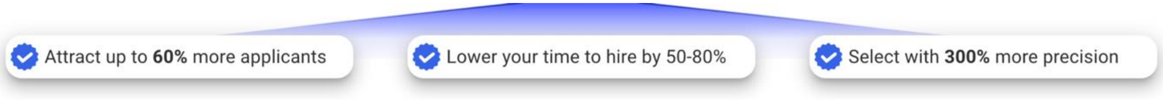


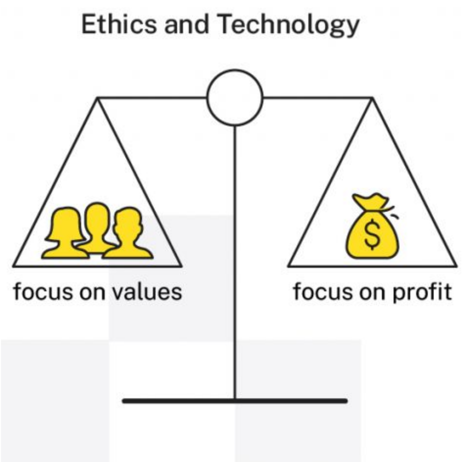



# VSD Perspective on Typical AI Candidate Ranking Systems



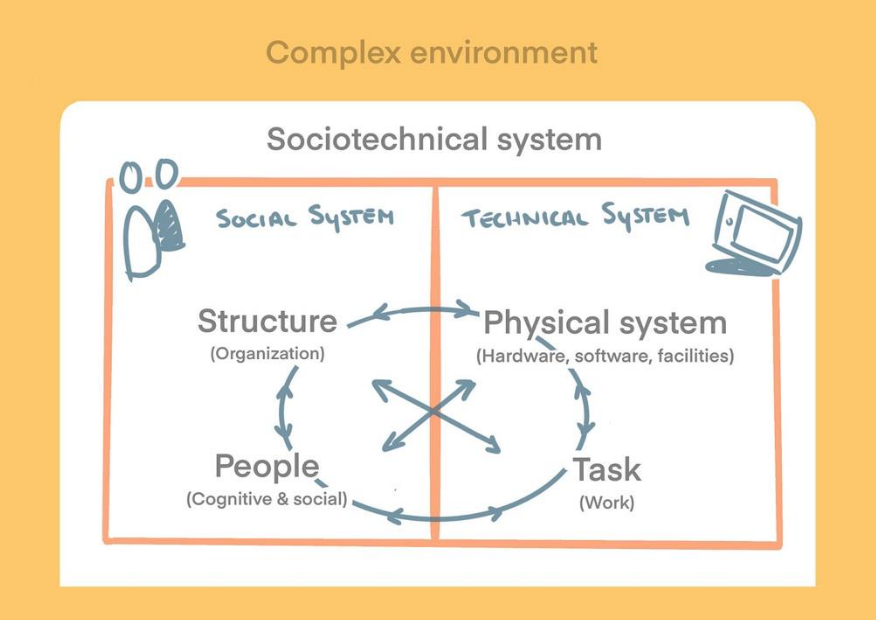

- *Values*
  - Profit, Efficiency, DEI, Accuracy, Scaling, Company Security.
- *Normative Assumptions*
  - Precision -> Assumption on ground truth (know ahead of time if the candidate is good, based on historical data).
  - Candidates can be ranked from best to worst (job suitability is 1-dimensional).
  - Saving time and money is always good.
  - More applicants is better.
- *Stakeholders addressed in design*
  - Tool producer, target user: companies/HR.





# Sociotechnical Context

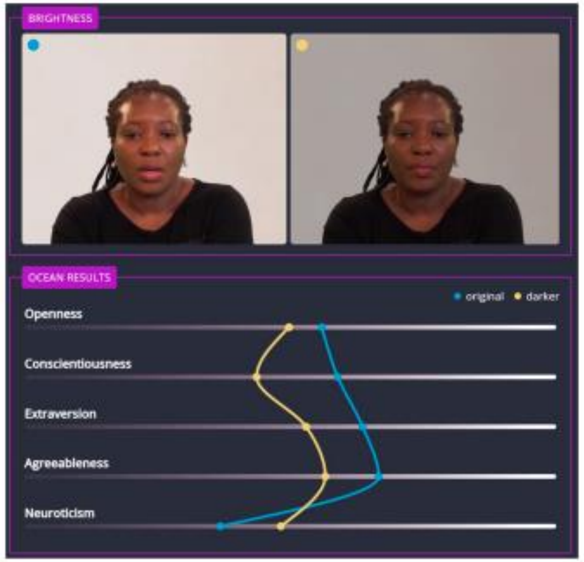
- What is the tool's context? What are the present dynamics and structures?
- Power imbalances
  - Many applicants, single job (amplified by LinkedIn).
  - Information/Transparency - Hiring practices, salary, workplace environment.
- Company profit-motive vs. social justice
  - Without financial incentive, social justice relies on policy, regulation, audits, enforcement…



# Values and Biases in the Data

- Pre-training
  - Large language models, Face analysis software
    - -> Model's Umwelt predominantly English and light-skinned.
    - Normative assumption that these are more important?
- Training
  - Assumption that historical data distribution should be reproduced (value status quo).
  - Which identities/domains/professions are neglected?
  - LinkedIn Engangement Gender Bias
- Measurement
  - Reduction of candidate to only the collected and processed data (anything else doesn't matter).
- Annotations
  - Values and biases of annotators.

| | 1st | 2nd | 3rd | 4th | 5th |
|---|---|---|---|---|---|
| **GPT-3 training data (2019)** [35] | English (93%) | French (1.8%), | German (1.5%) | Spanish (0.8%) | Italian (0.6%) |
| **Languages represented on the Internet (2021)** [36] | English (44.9%) | Russian (7.2%) | German (5.9%) | Chinese languages (4.6%) | Japanese (4.5%) |
| **First-languages spoken (2019)** [37] | Mandarin Chinese (12%) | Spanish (6%), | English (5%), | Hindi (4.4%), | Bengali (4%). |
| **Most spoken language (2021)**[37] | English (1348M) | Mandarin Chinese (1120M) | Hindi (600M) | Spanish (543M) | Standard Arabic (274M) |



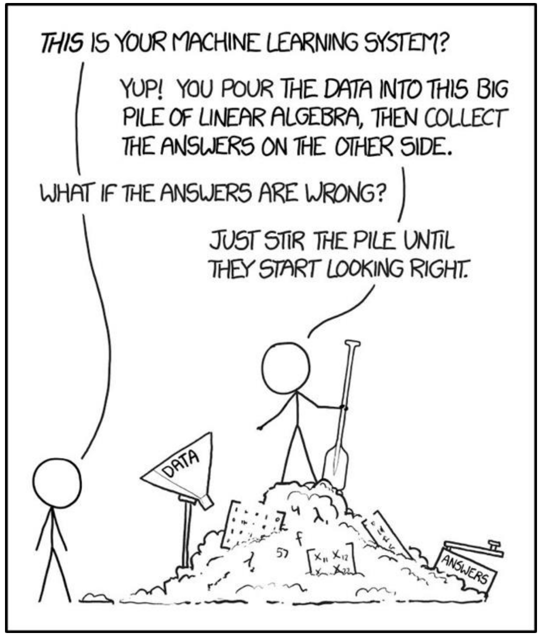
# Models - Values in the Math

- Model Architecture
  - *Ranking* – Job suitability is 1D; all candidates who don't meet certain threshold should be blindly discarded.
  - *Statistical outliers* – Out-of-box candidates should be ignored.
  - *Deep learning* – Accuracy over interpretability.
  - *Systematic error/bias* – No failsafe (e.g. via randomness in recommendations).
- Training
  - *Loss function* – Not neutral, but reinforces status quo.
  - Can make other value-based choices.
- Evaluation
  - *Accuracy* – Like loss, is not neutral.

The company might have historically hired fewer women and try to make up for the deficiency in training data with a loss function like:

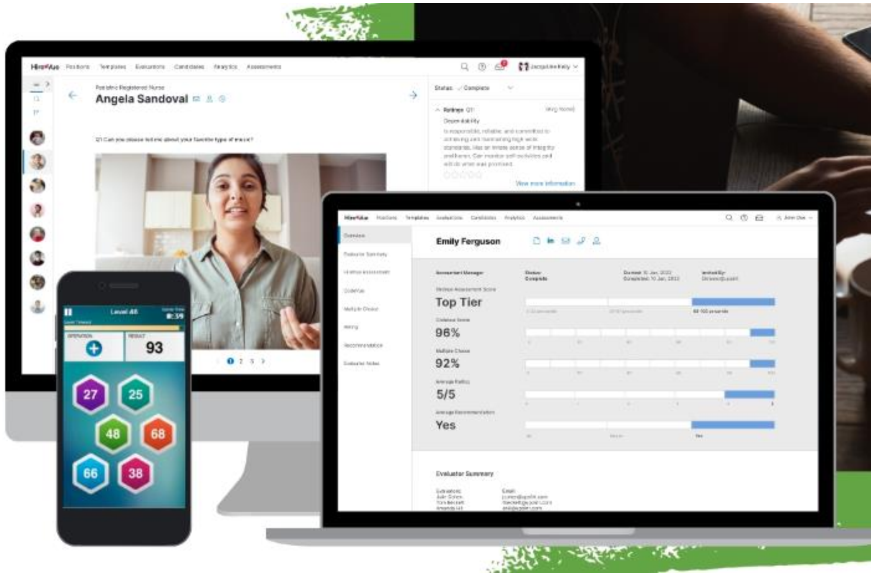
$$Loss = \frac{1}{n_{♂}} \sum_{i=1}^{n_{♂}} ((y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) + \frac{2}{n_{♀}} \sum_{i=1}^{n_{♀}} ((y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i))$$

I.e., misclassifications of female applicants are twice as bad as for men.
Of course, also assumes gender binary.



# Deployment

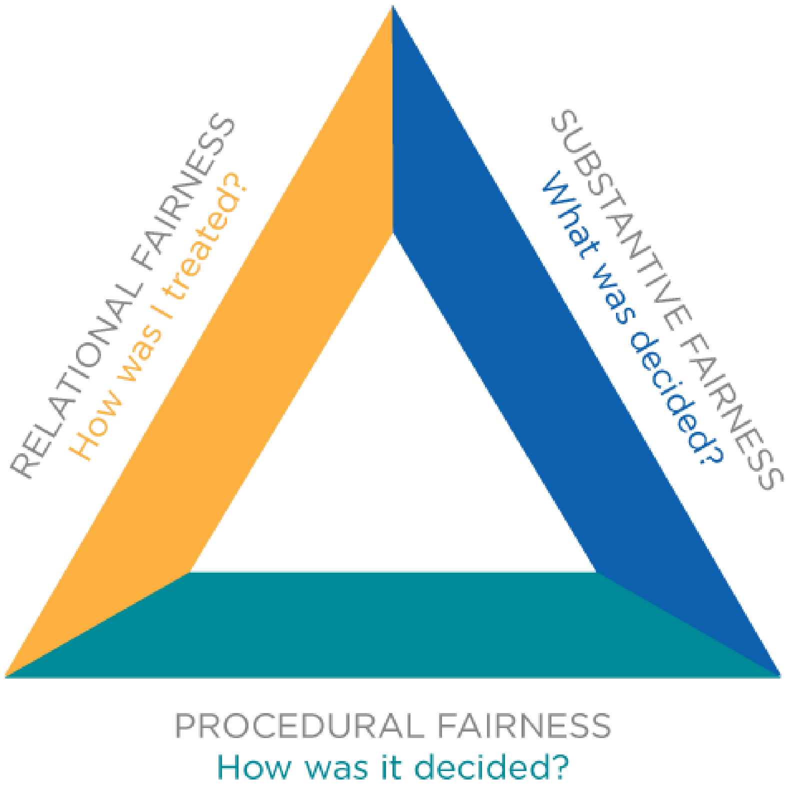
- Interface - All displayed info nudges user’s decision.
  - Just ranking and scores – Appearance of numerical objectivity.
  - Fixed recommendations or flexible?
  - Display Faces – Susceptible to tool user’s biases?
  - What isn’t displayed?
- Decision-Making
  - Pre-screening is usually automated. Assumption that computer is more precise and objective. No accountability.
  - Human-AI collab? Collective bias reinforced or mitigated?
- Feedback
  - Can user give feedback?
  - Does cand




# Should AI play a role in hiring decisions?

And if so, what?

# Investigating Fairness

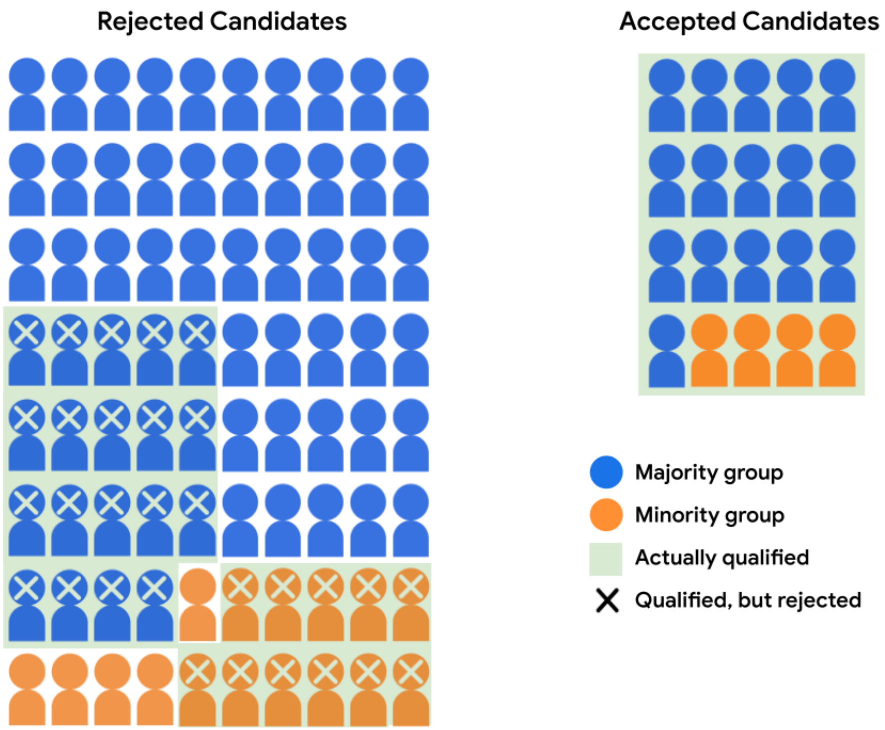


# Substantive Fairness

Most research on approaches to fairness in AI focus on substantive fairness on the level **model output with respect to different demographic groups**.

Different fairness metrics encode different normative assumptions ([Weerts et al. 2023](#))

- *Demographic parity*: The success rate of candidates should not depend on their background.
    - Doesn't use ground truth.
    - Doesn't account for under-representation of groups in applicant pool.
    - Might be interpreted as affirmative action or unfair penalization of historically preferred group.



# Substantive Fairness - Interventions

The normative underpinnings of methods for improving algorithmic (substantive) fairness should also be considered.

- *Counterfactual Data Augmentation* - E.g., add new successful minority candidate data samples by modifying samples from majority.
  - Increases preference for minority candidates that **most closely resemble the majority** (who may be very rare in reality).
  - Promotes diversity, but not inclusion.

Fairness benchmarks can be gamed without supporting real-world fairness.

- E.g., dropping rejected minority candidates from test data to achieve demographic parity.

**Upshot**: Technical fairness methods should align with the underlying ethical definition of fairness.

# Relational Fairness

For the most part, rejected candidates receive no personal feedback.

There's no way to tell how their data was processed, or even if it was processed correctly.

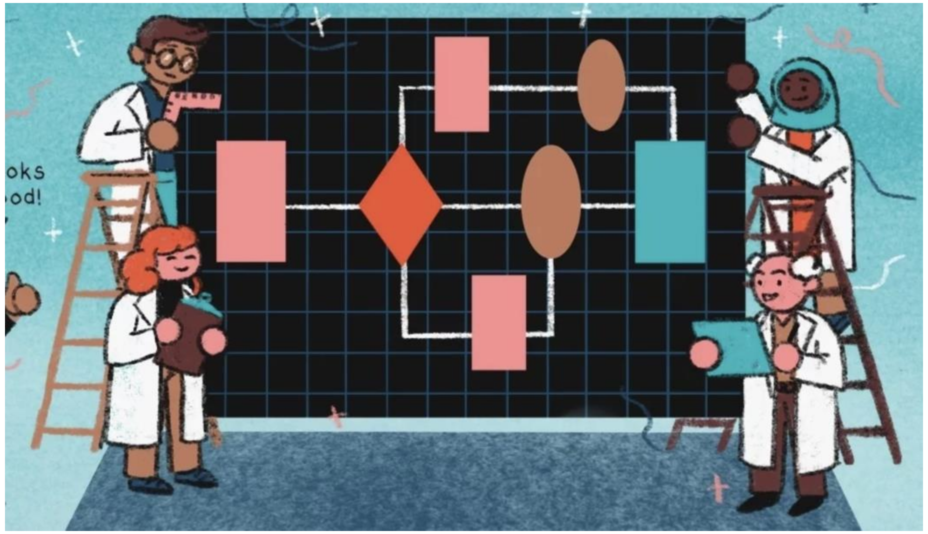
How to improve treatment of job applicants? Make the interaction feel more balanced and equal?

- *Agency*
  - Allow applicants to view post-processed model input data (e.g., to avoid rejections from simple parsing errors)?
  - Appeal to human if justified suspicion of algorithmic bias?
- *Justification*
  - Human/AI-generated personalized justifications for rejection?
- *Auditing*
  - Suggested auditing tips for applicants (lighting, hair, gender or ethnicity swap)?

# Procedural Fairness

How to gauge if the decision-making process is fair?

- *Explainable AI (XAI)* - Mostly limited to evaluating if the model used reasonable features from candidate data to render its decision.
- *Transparency* - E.g., using simpler alternative models with clearer decision-making process.
- *User-agency* - If the decision ultimately made by the tool's user…
  - Fairness in how interaction with AI affects human decision is more important.
  - Give user more direct influence over tool parameters/make company accountable for decisions.
  - More interactive (chat-based?) approach/generate justifications for recommended/rejected candidates to support user's decision-making.

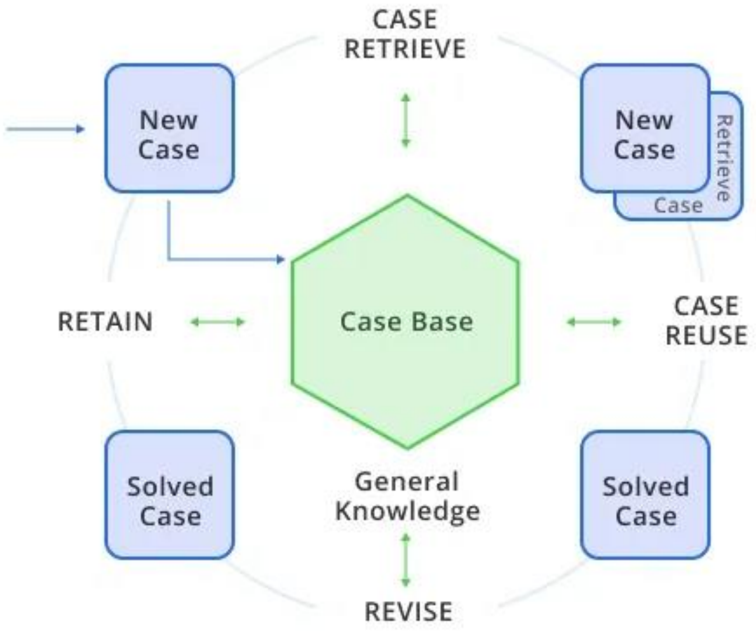


# Existing Approaches – A Summary

- The primary focus is on group fairness in terms of demographic distributions in outcomes.
- The normative assumptions (about fairness, bias, etc.) underlying metrics and technical interventions typically go unstated.
- This makes it difficult to know if metrics and methods are actually aligned with ethical values.
- Methods tweak existing models for harm mitigation and measure effectiveness via fairness metrics, rather than ground-up design from fairness stance.

# Technical Investigation – Proactive Design

# The Case-Based Reasoning Alternative



# Case-Based Reasoning (CBR)

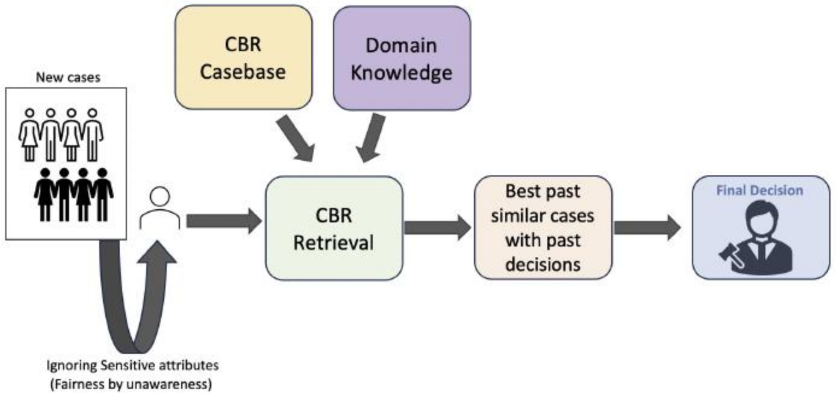



**Figure 1:** CBR Pipeline

$$S_{\text{global}}(X, Y) = \sum_{i=1}^{n} w_i \cdot s_{\text{local},i}(x_i, y_i), \tag{1}$$

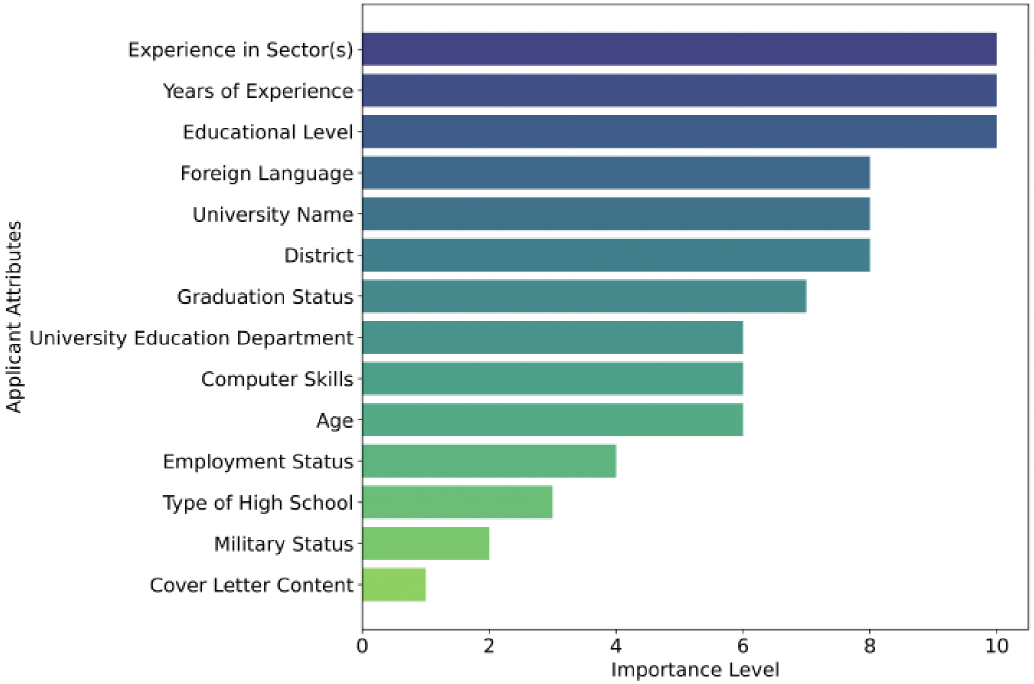



**Figure 2:** Survey results depicting the general importance of attributes for recruitment decisions.

# Operationalizing Fairness

- No direct information about protected attributes.
- Expert-informed feature selection and customizable weights.
- Principle "similar candidates treated similarly" (relates to consistency).
- Explainable. Technically robust.
- Certainty metric (based on distance to similar cases) related to procedural fairness and allows flagging of outliers.
- Case-base rapidly updated (easy to fix compared to ML).
- Shortlisting, not ranking.

# Parallel work…

- Proxy discrimination detection in data (e.g. protected attribute prediction not better than random).
- Context- and culture-specific bias detection and mitigation in (large) language models.
- Flagging grounds for potential unfair discrimination in candidate data.

# Future directions…

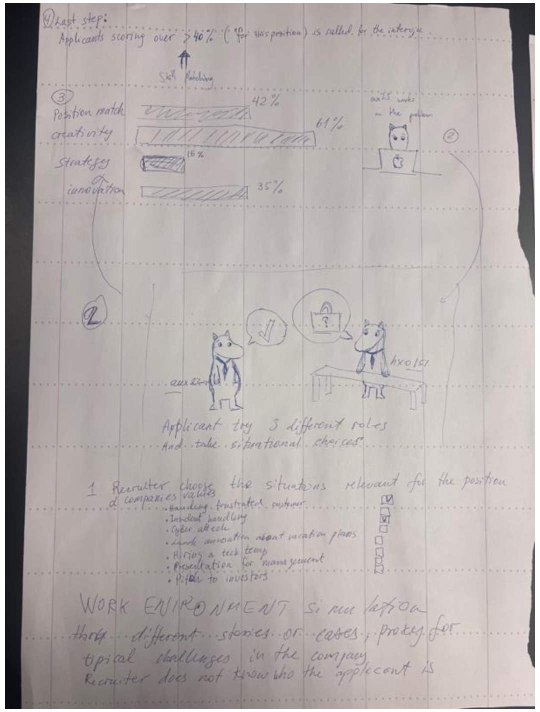
*Justifying Decisions* – combining XAI with LLM justifications.

*Agonistic ML* – The *incomputable self* ->

- Tools based on customizability, adaptability, ensemble models (e.g., CBR + blackbox NN + LLM), deliberative processes…
- Explicit combining and balancing of values.
- Chance for candidates to demonstrate/represent themselves in best way for given job.

**Choose the top 5 evaluation features based on the job ad**

- [ ] Perfect match (100/100) with must-have requirements
- [ ] Perfect match (100/100) with nice-to-have requirements
- [ ] Education in relevant field/discipline
- [x] Formal education level
  - (•) Must have ( ) Nice to have
- [ ] Relevant professional experiences
- [ ] Having professional experiences in a different field of interest
- [ ] Quantity of previous experiences (the more the better)
- [ ] Duration of previous experiences (the longer the better)
- [ ] Relevance ONLY of the latest work experience
- [x] Prestige of past experiences
  - (•) Must have ( ) Nice to have
- [ ] Meeting ONLY the hard skills
- [ ] Language Level requested in the job offer
- [ ] Having the required licenses
- [ ] Candidate matching Diversity & Inclusion policies' criteria
- [ ] Cultural and Value fit for the company
- [x] Availability to flexible working hours
  - (•) Must have ( ) Nice to have
- [ ] Distance from candidate and job's location
- [ ] Not being overqualified
- [ ] Law and policy compliance (visa requirements etc)
- [ ] Communication and social skills
- [ ] Different interesting hobbies
- [ ] Motivation and Resourcefulness
- [ ] Lack of experience compensated by a strong educational background




- [ ] Add a feature

**AI Ranking**

Candidate: Mohammed
Ranking: 1
Explanation: Mohammed has an extensive background in retail management, boasting over 20 years of experience as a Regional Manager at MegaMart Retail, which is a prestigious position in the industry. He has excellent leadership and management skills, as well as fluency in Dutch, English, and Arabic. Despite having a high school diploma, Mohammed's considerable industry experience qualifies him as an excellent match for the Assistant Store Manager position, especially in a flagship store setting. His willingness for flexible work hours makes him highly suitable for this role.

Candidate: Felix
Ranking: 2
Explanation: Felix holds a Bachelor's Degree in Economics and has relevant experience as an Assistant Store Manager and Logistics Officer, closely aligning with the job's requirements. Despite having relocated to Switzerland, he is available for flexible working hours. Felix's language skills include English and basic Dutch, which accommodate the role's location. His attention to detail and logistics experience is valuable for store operations. The lack of long-term experience compared to Mohammed is balanced by his solid educational background and relevant recent experience.

Candidate: Isabela
Ranking: 3
Explanation: Isabela has solid experience as a Store Manager and is pursuing an MA in Industrial Engineering, showing strong educational credentials. However, as they are currently on a student visa, their legal capacity to work full-time might be limited. They do offer flexibility with working hours. Although their previous experience is slightly less aligned with large-scale retail operations, their educational pursuit is a positive indicator for further growth and potential compensation for less direct experience.

Candidate: Priya
Ranking: 4
Explanation: Priya has relevant experience in store and quality management but lacks recent retail management experience. Her education level is a high school diploma, with further experience oriented towards customer support and food manufacturing quality. She has the flexibility needed for the position and some work experience in large-scale retail operations, but her recent roles don't closely align with an Assistant Store Manager position. However, her intermediate Dutch language skillset may require improvement for this Amsterdam-based job.